Institute of Paper Science and Technology
Atlanta, Georgia

IPST Technical Paper Series Number 935

Non-Linear Factor Analysis (NLFA) with Feedforward Networks Performs Non-Linear Data Reduction with Extraction of Linear Scores

S. Karrila

February 2002

Submitted to
ISDA 2002: Second International Workshop on
Intelligent Systems Design and Applications
Atlanta, Georgia
August 7-8, 2002

INSTITUTE OF PAPER SCIENCE AND TECHNOLOGY
PURPOSE AND MISSIONS

The Institute of Paper Science and Technology is an independent graduate school, research organization, and information center for science and technology mainly concerned with manufacture and uses of pulp, paper, paperboard, and other forest products and byproducts. Established in 1929 as the Institute of Paper Chemistry, the Institute provides research and information services to the wood, fiber, and allied industries in a unique partnership between education and business. The Institute is supported by 51 member companies. The purpose of the Institute is fulfilled through four missions, which are:

- to provide a multidisciplinary graduate education to students who advance the science and technology of the industry and who rise into leadership positions within the industry;
- to conduct and foster research that creates knowledge to satisfy the technological needs of the industry;
- to provide the information, expertise, and interactive learning that enable customers to improve job knowledge and business performance;
- to aggressively seek out technological opportunities and facilitate the transfer and implementation of those technologies in collaboration with industry partners.

ACCREDITATION

The Institute of Paper Science and Technology is accredited by the Commission on Colleges of the Southern Association of Colleges and Schools to award the Master of Science and Doctor of Philosophy degrees.

NOTICE AND DISCLAIMER

The Institute of Paper Science and Technology (IPST) has provided a high standard of professional service and has put forth its best efforts within the time and funds available for this project. The information and conclusions are advisory and are intended only for internal use by any company who may receive this report. Each company must decide for itself the best approach to solving any problems it may have and how, or whether, this reported information should be considered in its approach.

IPST does not recommend particular products, procedures, materials, or service. These are included only in the interest of completeness within a laboratory context and budgetary constraint. Actual products, materials, and services used may differ and are peculiar to the operations of each company.

In no event shall IPST or its employees and agents have any obligation or liability for damages including, but not limited to, consequential damages arising out of or in connection with any company's use of or inability to use the reported information. IPST provides no warranty or guaranty of results.

Non-Linear Factor Analysis (NLFA) with Feedforward Networks Performs Non-Linear Data Reduction with Extraction of Linear Scores

Seppo Karrila
Institute of Paper Science and Technology
Atlanta, GA 30318-5794

ABSTRACT

While Principal Component Analysis (PCA) is a well-known method for quantifying the redundancy between observed numerical values, it cannot effectively reduce non-linear relationships between these variables. Non-Linear Factor Analysis (NLFA) is a novel auto-associative method for data reduction, which extracts linear features to provide an easily documented encoding, while the decoding mapping is non-linear. NLFA can be implemented with standard feedforward networks if the user has control over the network configuration; many commercial packages provide the functionality needed.

This type of auto-association for data reduction has not been studied or reported earlier in any depth. The (feature selection –type) encoding-decoding pair $(x, y(x)) \mapsto x \mapsto (x, y(x))$, where x and y may be vectors and $y(\bullet)$ is a non-linear mapping, is a natural example of linear encoding whose inverse (decoding) is non-linear. Linear feature extraction with non-linear decoding can be viewed as a generalization of feature selection – a heuristic discussion suggests this type of data reduction is effective for a wide range of practical non-linear problems. The level of effectiveness is discussed through Linearly Reducible Intrinsic Dimensionality (LRID), which is defined for continuous error-free data in a manifold; the NLFA method estimates this value from representative (typically inaccurate) discrete data.

The numerically found linear encoding can be used to improve (post-process) the decoding in a manner that corresponds to a decoding network with bypass connections, and reduces the reconstruction error for the discrete data. This improved NLFA ensures that the composite function performing encoding followed by decoding, $p=g \circ f$, is idempotent, $p^2=p$, which ensures consistent results when the mapping p is viewed as "a projection to corrected values" that are in the range of the decoding g.

The data reduction capacity of NLFA is between conventional PCA and the fully non-linear Kramer's NLPCA; the relationship of these three methods is discussed. Reference is provided to a separate publication, which contains numerical application examples of the basic (non-improved) NLFA.

INTRODUCTION

Data reduction seeks to encode data vectors to a lesser apparent dimensionality, in such a manner that no essential information is lost. For application to new, previously unseen data, this reduction process needs to take the form of an encoding mapping that compacts a data vector, and a decoding mapping that reconstructs the original data vector from the encoded form.

Prior work using neural networks for data reduction has concentrated on encoding-decoding pairs with both mappings either linear or non-linear – it turns out that the combination of linear with non-linear is beneficial, as long as it is the encoding that is linear. The linear encoding can be easily documented as a matrix, post-processed to orthogonalize or rotate the "factor loadings", interpreted to guide feature selection, and used to improve the decoding mapping as will be shown. Linearly encoded normally distributed variables are also normally distributed, which is a benefit for further statistical analysis.

BACKGROUND

Gaussian Elimination and Feature Selection

Given a set of linear or non-linear algebraic equations, a natural approach is to proceed as with Gaussian elimination. If possible, the first equation is solved for a variable, the solution is substituted to the

remaining equations to eliminate this variable, and the remaining equation set is treated similarly – this process is reiterated until no further elimination is possible. A linear system would now be in upper triangular form, and conversion to diagonal form would proceed by back-substitution in reverse order – this back-substitution is also doable for non-linear equations. The final form of equations assigns values to the solved dependent variables, so that only the remaining independent variables (parameters) are used to compute the values. Fundamental physical equations are often such, that this approach works well with them. When some variables in a redundant set can be expressed as single-valued functions of the remaining variables, feature selection will be able to reduce the corresponding data.

Some theoretical results that pertain to such feature selection are now briefly discussed.

For linear equation systems global theorems basically state that the number of independent equations is the rank of the coefficient matrix, and this is the number of variables that we can eliminate from an underdetermined system that has a solution. The variables to be eliminated will be so selected that their coefficient matrix has full rank.

For non-linear equation systems local results analogous to "Gaussian elimination" are well known. These theorems in multivariable calculus or theory of differential forms are known as the Implicit Function Theorem, the Inverse Function Theorem, and the Rank Theorem, or they are discussed in connection with functional dependency. In a small neighborhood of an existing solution the non-linear functions behave almost linearly, and the rank of the Jacobian matrix determines how many variables can be eliminated – solved in terms of the remaining variables or parameters as unique single-valued functions. Feature selection is then guaranteed to work also with non-linear systems if the variable ranges are small enough.

In summary, even when multiple variables have non-linear constraints, it is often possible to select a subset of them, which determines the rest as single-valued functions. If this is not possible globally, the domain needs to be subdivided into small enough neighborhoods.

Manifolds and Intrinsic Dimensionality

Assume that some reasonably smooth model equations (fundamental, phenomenological, or empirical) constrain the measured variables. Mathematically, under some smoothness conditions and within an open connected set, a system of constraints $g(x)=0$, where g is a vector function having a Jacobian of constant rank r and x is n-dimensional, defines a k-dimensional manifold in $\mathbb{R}^n$, with $k=n-r$. This implies that k locally selected features or coordinates determine through smooth mappings all components of x, i.e., the data is locally reducible to k features and nothing less will do – the whole manifold may be a union of such local patches.

The intrinsic dimensionality of a discrete data set is typically defined along the lines: "the intrinsic dimensionality is the minimum number of independent variables needed, so that encoding to these variables and decoding back to the original variables is possible without significant loss of information". The discrete definition is of necessity vague, leaving room for interpretation. A definition that could be made rigorous, motivated by the discussion above for continuous error-free variables, is: "the intrinsic dimensionality of a continuum of data vectors x is the smallest such k that the vectors x are a subset of a k-dimensional manifold".

Definition of Linearly Reducible Intrinsic Dimensionality

The definition of intrinsic dimensionality for non-linear redundancies is based on local concepts. The objective of this section is to provide a global weaker concept that has significance for practical data analysis. The concept will not be based on feature selection, because then it would be dependent on the coordinate system used.

Feature selection is equivalent to specific orthogonal projections– those that eliminate some of the coordinates. Just as projections are a special case of linear mappings, data reduction by feature selection is a special case of linear feature extraction – the latter has better reduction capacity in terms of remaining dimensionality. Further, neural networks can find optimal linear mappings by optimizing their weights so

it is prudent to define the extent of data reduction possible by this approach, as regards continuous error-free data in a manifold.

Definition of LRID. A manifold M in $\mathbb{R}^n$ is linearly reducible to dimension m, if there is a linear mapping A into $\mathbb{R}^m$ such that the restriction of A to M is bijective (= one-to-one). The linearly reducible intrinsic dimensionality (LRID) of M is the smallest of such values m.

By allowing linear mappings we ensure that the LRID is independent of rotation, translation, stretching, or shearing of M. This makes the LRID an intrinsic property of the manifold M, independent of how the observer positions his linear, not necessarily orthogonal, coordinate axes.

If $m<n$, then data reduction is performed by the encoding-decoding pair $x \mapsto Ax \mapsto x$. Denoting the (non-linear) decoding by g, $x=g(Ax)$ for all x in M. If m=LRID, then A must be onto $\mathbb{R}^m$ and of full row rank, and AA^* is invertible – this is needed to improve g.

On Estimating the Intrinsic Dimensionality of Discrete Data

Defining intrinsic dimensionality based on manifold concepts requires a continuum of data. Connectionist or statistical approaches deal with finite discrete data sets, and a reasonable interpretation is needed to bridge the continuous models and theory with discrete data.

Consider a finite set of points in the x-y plane. The superficial dimensionality (number of variables in each data record) of this data is two, but if the points were visually neat "on a nice curve", an algorithm estimating the intrinsic dimensionality should return value one.

Even with completely random points only a finite number of lines join pairs of these points, so we can choose a projection direction not parallel to any of these lines. Then this projection gives a different image P(x,y) for each of the data points (x,y). An interpolating function that maps each P(x,y) to (x,y) provides a curve through the data. Then it would seem that any finite set of data points in the plane is actually one-dimensional in the sense of LRID.

Mathematically this is quite correct, it is the predictive ability of the result that must be questioned – the perceived problem is closely related to the concept of over-fitting neural networks. The purpose of joining points in a plane by a curve is to show an interpolation between the points. This interpolation must be tested, as is conventionally done with neural networks (NN) that are trained with one data set and validated with another. Constructive search for the LRID using an NN method then automatically deals with the recognized problem of intrinsic dimensionality definition for discrete data. The user is still left with the judgment call of what level of reconstruction error is acceptable, but this is the case also with well-understood linear PCA – we cannot expect to do better with a non-linear reduction method.

Prior Neural Network Methods for Data Reduction

Principal Component Analysis (PCA) is a well-established method, mathematically related to the Singular Value Decomposition or Polar Decomposition of a matrix. It finds the unique affine subspace of given dimensionality, such that the orthogonal projection of data into this subspace retains a maximal fraction of the variance; the extracted features are orthogonal projections to the axes spanning this subspace. Factor Analysis can be considered post-processing of the PCA results, by selecting a new, possibly oblique, set of axes (approximately) spanning the same subspace, often so that each of the feature values (called scores) is referred back to a small subset of the original variables.

A feedforward neural network processes a data vector sequentially layer by layer. The outputs of any layer can be viewed as encoded values, mapped to the outputs of the network by the remaining layers. If the output target values are equal to the inputs, the network is auto-associative. Then the encoded values are approximately mapped back to the inputs – the same network also embeds the decoding mapping in its structure.

If an auto-associative neural network (AANN) is successfully trained, it has learned encoding-decoding mappings that preserve the original data vectors with only a small reconstruction error (and

when a validation set is used, the ability to interpolate has also been tested during training and network selection). If the AANN has a bottleneck layer with a small number of nodes, the encoding at this layer reduces the dimensionality of the data vectors. This is how the bottleneck AANN structure functions as a tool for data reduction.

An AANN with linear activation functions and only one hidden bottleneck layer performs PCA. It finds the same subspace as PCA would find for a given reduction of dimensionality, but the factor loadings (weight vectors of linear encoding) will not be orthogonal without post-processing or special network constructs. An extensive review is provided in a recent book (Diamantaras and Kung, 1996).

Kramer's non-linear PCA (NLPCA) learns non-linear mappings f and g to perform encoding and decoding. These mappings are each represented by an NN with (at least) one hidden sigmoidal layer, and when the two networks are combined at the bottleneck so that the output layer of f is the input layer of g, the AANN has (at least) three hidden layers. The universal approximation property of NN ensures that three hidden layers with enough nodes in the first and third will always suffice, and this is the configuration that Kramer originally presented (Kramer, 1991). NN training becomes more difficult as network depth is increased and even with convergence the weights may be stuck to a local suboptimal error minimum – this encourages to limit the network depth, but on occasions increasing the number of layers is useful.

An invertible mapping applied to the reduced values can be used to form new encoding-decoding pairs, so this non-linear reduction is not unique (or user independent) like linear PCA is. While Kramer discusses the application of information theoretic principles to select the reduced dimensionality, using the validation set seems to be a good practical approach for avoiding over-fitting also with AANN.

THE NLFA METHOD

The NLFA is a constructive NN-method to estimate the LRID and to provide the corresponding mappings, using conventional feedforward networks of the bottleneck AANN -type.

Let the manifold M correspond to the continuous model represented by discrete data. According to the definition of LRID a linear map A of full row rank exists, such that $x=g(Ax)$ is a decomposition of the identity mapping on M – an auto-associative identity that can be attempted with an AANN and discrete data. The manifold M is (a connected part of) the range of g in $\mathbb{R}^n$, and new data points can be mapped into M by p: $x \mapsto g(Ax)$. Since the same mapping is just identity on M, repeated application of p will not change the image: $p=p\circ p$ or $p(\bullet)=p(p(\bullet))$. This idempotent mapping p is termed projection to corrected values since it maps a new data point so that the constraints intrinsic to the manifold M are satisfied. (Lowercase p is used to indicate that this is not a linear mapping.) The AANN provides numerical approximations A and g, and the estimate of p constructed from these will only be close to idempotent in the range of g – a method of correction will be provided for NLFA to fix this problem.

Kramer's NLPCA method has been criticized because its projection gof to corrected values can behave erratically with the compounding non-linearity (Malthouse, 1998). The projection p of NLFA maps x into M along the orthogonal complement of the row space of A because $p(x')=p(x) \Leftrightarrow Ax'=Ax \Leftrightarrow \{x'-x$ is orthogonal to all the rows of $A\}$; the constant value surfaces of p are affine manifolds along which a new point is "corrected", i.e., mapped into M.

In analogy with Principal Component Analysis (PCA) or Factor Analysis, the components of Ax can be called scores, while the rows of A represent factor loadings.

Implementation

The NLFA is implemented with an AANN whose first hidden layer is a bottleneck with linear activations; the outputs from the bottleneck provide the encoded values Ax. At least one more hidden layer with sigmoidal activations is needed for the non-linear decoding g, so minimally a two-hidden-layer NN is required to perform NLFA.

As with Kramer's NLPCA, various numbers of nodes in the bottleneck need to be tried to find the smallest number that provides acceptable reconstruction error. Even with a single network configuration it

is necessary to reiterate the training with various initializations to avoid worst cases of convergence to a local (non-global) optimum.

Improvement of decoding

Given an approximate encoding-decoding pair, the decoding defines the manifold by M=range(g). As the encoding f is not exactly the inverse of g on M, p=g∘f is not a decomposition of identity on M. Repeated application of p will possibly cause "creep" along M, and p can not be used as a "projection to corrected values". A numerical minimization procedure could be implemented based on mapping g alone, but such iterations can be avoided with the NLFA.

In accordance with the definition of LRID, f=A is of full row rank. The orthogonal projection $Q=A^*(AA^*)^{-1}A$ is such that

$$AQ=A \tag{1}$$
$$A(1-Q)=AQ(1-Q)=A0=0 \tag{2}$$
$$(1-Q)A^*=0 \tag{3}$$

Define an improved decoding by

$$h=(1-Q)g + A^*(AA^*)^{-1} \tag{4}$$

Then for p=h∘A it is easy to show that pQ=p and Qp=Q from which

$$p(p(x))=p(Qp(x))=p(Qx)=p(x) \tag{5}$$

so p is everywhere idempotent, and identity on M=range(p)=range(h).

Further, for any x we have Qp(x)=Qx so all of the reconstruction error is along projection (1-Q). But (1-Q)p(x)=(1-Q)g(Ax) and this part of the reconstruction error is unchanged. By the Pythagorean theorem p(x) is a better approximation to x than the original reconstruction g(Ax). In particular the reconstruction error for the training and validation data sets is reduced.

Improvement of g to h both reduces the error sum with any data, and ensures that the encoding-decoding pair becomes idempotent. Note that the corrections done to g are a combination of linear post-processing and a linear bypass from input to output – they can be implemented also by including bypass connections and adjusting the output layer weights.

Comparison with Other Constructive Methods

Going backwards from NLPCA, if we restrict the encoding to be linear we have NLFA, and if we further restrict also the decoding to be linear we have PCA. Clearly the capacity to reduce data decreases with each restriction, so NLFA is a compromise in complexity and capacity between the two earlier methods – a semi-linear method between fully linear and fully non-linear.

To visualize a comparison between these methods, Figure 1 provides a taxonomy, which refines the conventional classification of data reduction methods to linear and non-linear. It is appropriate to categorize methods based on the encoding and decoding types separately. With the addition of the NLFA method to the arsenal, an approach exists for every useful slot in this taxonomy.

Note that if the decoding is linear and of full rank relative to its inputs, the generalized inverse of this mapping provides the encoding – non-linear encoding is useless in the first row of the tabulation. Linearity and non-linearity need to be combined in the right way to achieve a useful result.

The generalized inverse of a linear mapping B with full column rank is very similar to the second term on the right-hand-side of equation (4) above, and familiar from linear least squares – it is given by $(B^*B)^{-1}B^*$. Once a linear decoding B is known, it is easy to write an explicit formula for the encoding.

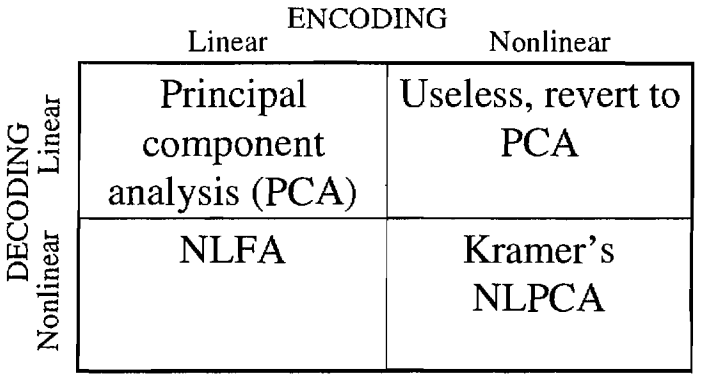

Figure 1. Types of encoding and decoding are used to create taxonomy of some constructive data reduction methods. NLFA stands for Non-Linear Factor Analysis, while NLPCA refers to Non-Linear PCA with neural networks.

CONCLUSIONS

A novel constructive method, the Non-Linear Factor Analysis (NLFA), performs data reduction by linear encoding and non-linear decoding. Such restricted data reduction will not necessarily find the intrinsic dimensionality of data; instead it can estimate the Linearly Reducible Intrinsic Dimensionality (LRID), which is defined based on continuous model equations that constrain the variables represented by the data.

NLFA has a better data reduction capacity than conventional PCA, but it is not as powerful as the fully non-linear Kramer's NLPCA. However, NLFA has benefits over NLPCA, which include:

- Linear encoding preserves normal error distributions and is easily documented.
- Relatively low complexity contributes to fast training and good convergence.
- Approximate projection to corrected values is well behaved.
- Approximate projection to corrected values can be made idempotent with simultaneous improvement in reconstruction accuracy by linear post-processing of the encoding and decoding mappings.

Due to restrictions on space, numerical examples are not included – some are available in a separate publication (Karrila and Rezak, 2002).

REFERENCES

1. Diamantaras, K.I., and Kung, S.Y. (1996). Principal Component Neural Networks: Theory and Applications, NY: Wiley, ISBN 0-471-05436-4.

2. Kramer, M.A. (1991). Nonlinear principal component analysis using autoassociative neural networks, AIChE J., vol. 37, no. 2, pp. 233-243.

3. Malthouse, E.C. (1998). Limitations of Nonlinear PCA as Performed with Generic Neural Networks, IEEE Transactions on Neural Networks, vol. 9, no. 1, pp. 165-173.

4. Karrila, S.J. and Rezak, S. (2002). Review, Developments and Pulp and Paper Research Applications of Data Reduction with Neural Networks, accepted for publication in Proceedings of TAPPI Paper Summit. Atlanta. GA. March 4-6. 2002.